Reproducible Research: Peer Assessment 2

Douglas Wirtz

January 29, 2016

Project Title

Severe Weather Impact on the Public Health and Economy in the US

Synopsis

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

Upon completion of the analysis, I have determined that tornadoes have the greatest impact when it comes to population health and floods have the greatest impact when it comes to economic consequences.

Data

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can download the file from the course web site:

- Dataset: Storm Data [47Mb]

There is also some documentation of the database available. Here you will find how some of the variables are constructed/defined.

- National Weather Service: Storm Data Documentation
- National Climatic Data Center Storms Events FAQ

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

Questions

The step-by-step analysis below addresses the following questions:

1. Across the United States, which types of events (as indicated in the **EVTYPE** variable) are most harmful with respect to population health?

2. Across the United States, which types of events have the greatest economic consequences?

Data Processing

Create directory and download data from NOAA Storm Database if they do not already exist. Then read data into R.

```
# if directory does not exist, then create the directory
if(!file.exists("data")){
        dir.create("data")
}
# if file does not exist in the directory, then download the file
if(!file.exists("data/repdata-data-StormData.csv.bz2")){

download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormDa
ta.csv.bz2", destfile = "data/repdata-data-StormData.csv.bz2")
}
#read in the data
data <- read.csv("data/repdata-data-StormData.csv.bz2")
```

Load the packages needed for data analysis.

```
library(ggplot2)
library(plyr)
```

Print the first 3 lines. The purpose is to figure out which variables will be important for answering the questions and to determine the relationships between the variables.

```
data[1:3,]

##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1 4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1 4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1 2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
```

```
## 2         NA         0                        2.0   150 2   0          0
## 3         NA         0                        0.1   123 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
```

Subset the data with only the variables needed to answer the questions.

```
data_sub <- subset(data, select = c("EVTYPE", "FATALITIES", "INJURIES",
                                    "PROPDMG", "PROPDMGEXP", "CROPDMG",
                                    "CROPDMGEXP"))
head(data_sub)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0          K       0
## 2 TORNADO          0        0     2.5          K       0
## 3 TORNADO          0        2    25.0          K       0
## 4 TORNADO          0        2     2.5          K       0
## 5 TORNADO          0        2     2.5          K       0
## 6 TORNADO          0        6     2.5          K       0
```

Events Most Harmful To Population Health

From the subset data, aggregate the sum of the fatalities and injuries based on event type.

```
data_health <- aggregate(cbind(FATALITIES, INJURIES) ~ EVTYPE, data_sub, sum)
```

Remove the event types with no fatalities or injuries, add fatalities and injuries together, and order data frame based on highest total.

```
# remove the entries with no fatalities or injuries
data_health <- data_health[data_health$FATALITIES > 0 | data_health$INJURIES
> 0,]
# add the fatalities and injuries together
data_health$TOTAL <- data_health$FATALITIES + data_health$INJURIES
# order the data frame with the highest total first
data_health <- data_health[order(data_health$TOTAL, decreasing = TRUE),]
data_health[1:10,]
```

```
##                EVTYPE FATALITIES INJURIES TOTAL
## 834           TORNADO       5633    91346 96979
## 130    EXCESSIVE HEAT       1903     6525  8428
## 856         TSTM WIND        504     6957  7461
## 170             FLOOD        470     6789  7259
## 464         LIGHTNING        816     5230  6046
## 275              HEAT        937     2100  3037
## 153       FLASH FLOOD        978     1777  2755
```

```
## 427          ICE STORM         89      1975  2064
## 760 THUNDERSTORM WIND        133      1488  1621
## 972       WINTER STORM        206      1321  1527
```

Events With The Greatest Economic Consequences

According to the Documentation, property and crop damage (`PROPDMG` and `CROPDMG`) are expanded using the characters in `PROPDMGEXP` and `CROPDMGEXP`. "K" = Thousands (10^3), "M" = Millions (10^6), and "B" = Billions (10^9).

Remove the entries in the subset data where `PROPDMGEXP` or `CROPDMGEXP` is not equal to one of the characters above.

```r
# make all the characters uppercase
data_sub$PROPDMGEXP <- toupper(data_sub$PROPDMGEXP)
data_sub$CROPDMGEXP <- toupper(data_sub$CROPDMGEXP)
# extract data with all rows containg "K", "M", or "B"
data_econ <- data_sub[data_sub$PROPDMGEXP == "K" | data_sub$CROPDMGEXP == "K"
|
                              data_sub$PROPDMGEXP == "M" |
data_sub$CROPDMGEXP == "M" |
                              data_sub$PROPDMGEXP == "B" |
data_sub$CROPDMGEXP == "B",]
# get all the variables in PROPDMGEXP and CROPDMGEXP
count(data_econ$PROPDMGEXP)
```

```
##   x   freq
## 1     4312
## 2 0      5
## 3 3      1
## 4 5      2
## 5 B     40
## 6 K 424665
## 7 M  11337
```

```r
count(data_econ$CROPDMGEXP)
```

```
##   x   freq
## 1   156484
## 2 ?      5
## 3 0     16
## 4 B      9
## 5 K 281853
## 6 M   1995
```

Adjust the subset data to replace the characters with numeric values.

```r
# create key and replace characters with numeric values for PROPDMGEXP
propkey <- c("\"\"" = 10^0, "0" = 10^0, "3" = 10^0, "5" = 10^0, "B" = 10^9,
"K" = 10^3, "M" = 10^6)
data_econ$PROPDMGEXP <- propkey[as.character(data_econ$PROPDMGEXP)]
data_econ$PROPDMGEXP[is.na(data_econ$PROPDMGEXP)] <- 10^0
```

```
# create key and replace characters with numeric values for CROPDMGEXP
cropkey <- c("\"\"" = 10^0, "?" = 10^0, "0" = 10^0, "B" = 10^9, "K" = 10^3,
"M" = 10^6)
data_econ$CROPDMGEXP <- cropkey[as.character(data_econ$CROPDMGEXP)]
data_econ$CROPDMGEXP[is.na(data_econ$CROPDMGEXP)] <- 10^0
```

Create two additional columns that combine the property and crop damage components

```
# multiply PROPDMG and PROPDMGEXP to create a new column PROPCOST
data_econ$PROPCOST <- data_econ$PROPDMG * data_econ$PROPDMGEXP
# multiply CROPDMG and CROPDMGEXP to create a new column CROPCOST
data_econ$CROPCOST <- data_econ$CROPDMG * data_econ$CROPDMGEXP
head(data_econ)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0       1000       0          1
## 2 TORNADO          0        0     2.5       1000       0          1
## 3 TORNADO          0        2    25.0       1000       0          1
## 4 TORNADO          0        2     2.5       1000       0          1
## 5 TORNADO          0        2     2.5       1000       0          1
## 6 TORNADO          0        6     2.5       1000       0          1
##   PROPCOST CROPCOST
## 1    25000        0
## 2     2500        0
## 3    25000        0
## 4     2500        0
## 5     2500        0
## 6     2500        0
```

From the new data, aggregate the sum of the property and crop damage based on event type.

```
data_econ <- aggregate(cbind(PROPCOST, CROPCOST) ~ EVTYPE, data_econ, sum)
```

Add the `PROPCOST` and `CROPCOST` together and order the data frame based on highest total

```
# add the PROPCOST and CROPCOST together
data_econ$TOTAL <- data_econ$PROPCOST + data_econ$CROPCOST
# order the data frame with the highest total first
data_econ <- data_econ[order(data_econ$TOTAL, decreasing = TRUE),]
data_econ[1:10,]
```

```
##                EVTYPE     PROPCOST    CROPCOST        TOTAL
## 70              FLOOD 144657709800  5661968450 150319678250
## 193 HURRICANE/TYPHOON  69305840000  2607872800  71913712800
## 351           TORNADO  56937160533   414953270  57352113803
## 297       STORM SURGE  43323536000        5000  43323541000
## 113              HAIL  15732266753  3025954453  18758221206
## 58        FLASH FLOOD  16140811638  1421317100  17562128738
## 38            DROUGHT   1046106000 13972566000  15018672000
## 185         HURRICANE  11868319010  2741910000  14610229010
```

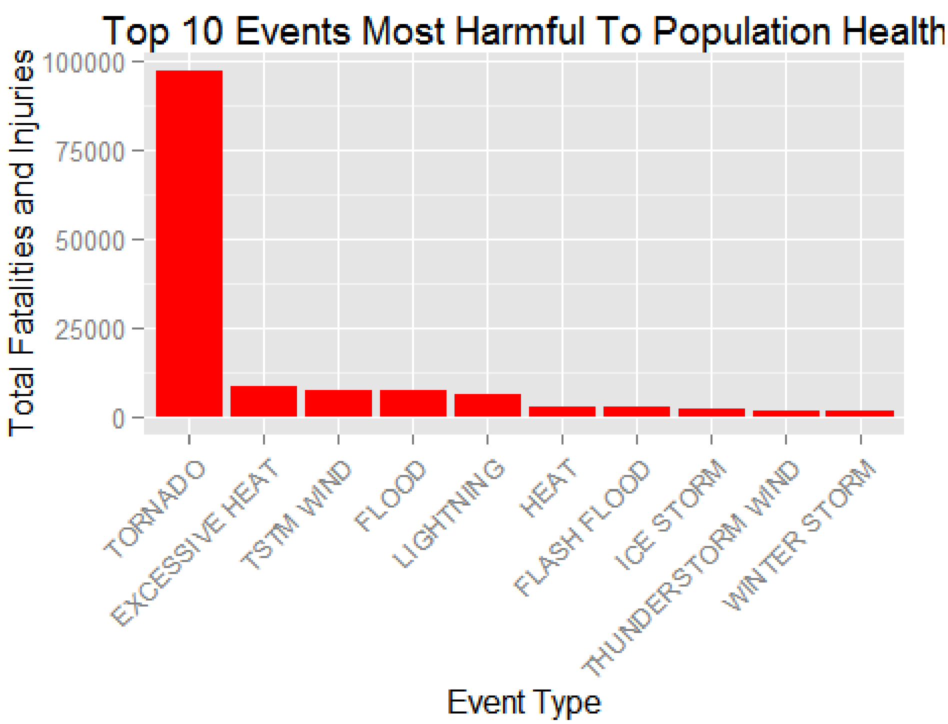
```
## 259       RIVER FLOOD   5118945500  5029459000  10148404500
## 202         ICE STORM   3944927860  5022113500   8967041360
```

Results

Top 10 Events Most Harmful To Population Health

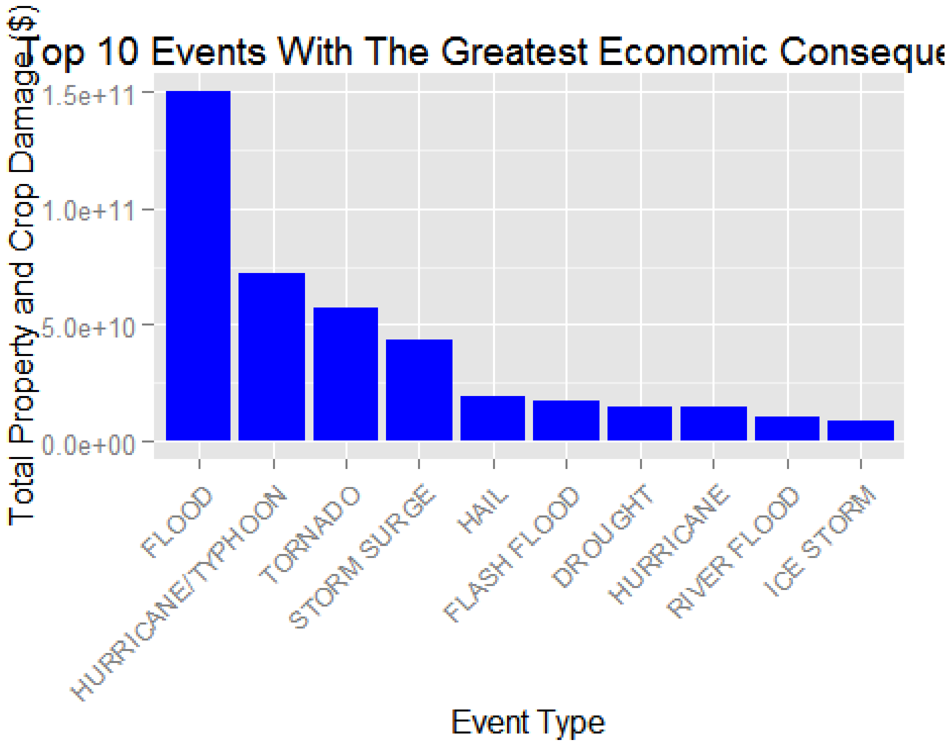
Graph the data to display which events had the greatest impact on fatalities and injuries.

```
ggplot(head(data_health, 10), aes(reorder(EVTYPE, -TOTAL), TOTAL)) +
        geom_bar(stat = "identity", fill = "red") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        xlab("Event Type") +
        ylab("Total Fatalities and Injuries") +
        ggtitle("Top 10 Events Most Harmful To Population Health")
```



Top 10 Events With The Greatest Economic Consequences Graph the data to display which events had the greatest impact on property and crop damage.

```
ggplot(head(data_econ, 10), aes(reorder(EVTYPE, -TOTAL), TOTAL)) +
        geom_bar(stat = "identity", fill = "blue") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        xlab("Event Type") +
        ylab("Total Property and Crop Damage ($)") +
        ggtitle("Top 10 Events With The Greatest Economic Consequences")
```



***

Conclusions

1. With respect to population health, tornado events had the greatest impact. This conclusion was based on the total number of fatalities and injuries in the US from 1950 to 2011.

2. With respect to economic consequences, flood events had the greatest impact. This conclusion was based on the total cost of property and crop damage in the US from 1950 to 2011.
